

Hash Function Learning via Codewords

Yinjie Huang, Michael Georgopoulos and Georgios C. Anagnostopoulos

CL, MG: EE & CS Dept., University of Central Florida; GCA: ECE Dept., Florida Institute of Technology

yhuang@eeecs.ucf.edu, michaelg@ucf.edu and georgio@fit.edu

Abstract

In this paper we introduce a novel hash learning framework that has two main distinguishing features, when compared to past approaches. First, it utilizes codewords in the Hamming space as ancillary means to accomplish its hash learning task. These codewords, which are inferred from the data, attempt to capture similarity aspects of the data's hash codes. Secondly and more importantly, the same framework is capable of addressing supervised, unsupervised and, even, semi-supervised hash learning tasks in a natural manner. A series of comparative experiments focused on content-based image retrieval highlights its performance advantages.¹

Keywords: Hash Function Learning, Codeword, Support Vector Machine

1 Introduction

With the explosive growth of web data including documents, images and videos, content-based image retrieval (CBIR) has attracted plenty of attention over the past years [3]. Given a query sample, a typical CBIR scheme retrieves samples stored in a database that are most similar to the query sample. The similarity is gauged in terms of a pre-specified distance metric and the retrieved samples are the nearest neighbors of the query point w.r.t. this metric. However, exhaustively comparing the query sample with every other sample in the database may be computationally expensive in many current practical settings. Additionally, most CBIR approaches may be hindered by the sheer size of each sample; for example, visual descriptors of an image or a video may number in the thousands. Furthermore, storage of these high-dimensional data also presents a challenge.

Considerable effort has been invested in designing hash functions transforming the original data into compact binary codes to reap the benefits of a potentially fast similarity search; note that hash functions are typically designed to preserve certain similarity qualities between the data. For example, approximate nearest neighbors (ANN) search [22] using compact binary codes in Hamming space was shown to achieve sub-linear searching time. Storage of the binary code is, obviously, also much more efficient.

Existing hashing methods can be divided into two categories: *data-independent* and *data-dependent*. The former category does not use a data-driven approach to choose the hash function. For example, Locality Sensitive Hashing (LSH) [4] randomly projects and thresholds data into the Hamming space for generating binary codes, where closely located (in terms of Euclidean distances in the data's native space) samples are likely to have similar binary codes. Furthermore, in [9], the authors proposed a method for ANN search using a learned Mahalanobis metric combined with LSH.

On the other hand, *data-dependent methods* can, in turn, be grouped into supervised, unsupervised and semi-supervised learning paradigms. The bulk of work in data-dependent hashing methods has been performed so far following the supervised learning paradigm. Recent work includes the Semantic Hashing [18], which designs the hash function using a Restricted Boltzmann Machine (RBM). Binary Reconstructive

¹This work has been accepted by ECML/PKDD 2015. Please cite the ECML version of this paper.

Embedding (BRE) in [10] tries to minimize a cost function measuring the difference between the original metric distances and the reconstructed distances in the Hamming space. Minimal Loss Hashing (MLH) [17] learns the hash function from pair-wise side information and the problem is formulated based on a bound inspired by the theory of structural Support Vector Machines [27]. In [16], a scenario is addressed, where a small portion of sample pairs are manually labeled as similar or dissimilar and proposes the Label-regularized Max-margin Partition algorithm. Moreover, Self-Taught Hashing [28] first identifies binary codes for given documents via unsupervised learning; next, classifiers are trained to predict codes for query documents. Additionally, Fisher Linear Discriminant Analysis (LDA) is employed in [21] to embed the original data to a lower dimensional space and hash codes are obtained subsequently via thresholding. Also, Boosting based Hashing is used in [20] and [1], in which a set of weak hash functions are learned according to the boosting framework. In [11], the hash functions are learned from triplets of side information; their method is designed to preserve the relative relationship reflected by the triplets and is optimized using column generation. Finally, Kernel Supervised Hashing (KSH) [13] introduces a kernel-based hashing method, which seems to exhibit remarkable experimental results.

As for unsupervised learning, several approaches have been proposed: Spectral Hashing (SPH) [26] designs the hash function by using spectral graph analysis with the assumption of a uniform data distribution. [14] proposed Anchor Graph Hashing (AGH). AGH uses a small-size anchor graph to approximate low-rank adjacency matrices that leads to computational savings. Also, in [5], the authors introduce Iterative Quantization, which tries to learn an orthogonal rotation matrix so that the quantization error of mapping the data to the vertices of the binary hypercube is minimized.

To the best of our knowledge, the only approach to date following a semi-supervised learning paradigm is Semi-Supervised Hashing (SSH) [25] [24]. The SSH framework minimizes an empirical error using labeled data, but to avoid over-fitting, its model also includes an information theoretic regularizer that utilizes both labeled and unlabeled data.

In this paper we propose *Supervised Hash Learning (*SHL) (* stands for all three learning paradigms), a novel hash function learning approach, which sets itself apart from past approaches in two major ways. First, it uses a set of Hamming space codewords that are learned during training in order to capture the intrinsic similarities between the data's hash codes, so that same-class data are grouped together. Unlabeled data also contribute to the adjustment of codewords leveraging from the inter-sample dissimilarities of their generated hash codes as measured by the Hamming metric. Due to these codeword-specific characteristics, a major advantage offered by *SHL is that it can naturally engage supervised, unsupervised and, even, semi-supervised hash learning tasks using a single formulation. Obviously, the latter ability readily allows *SHL to perform transductive hash learning.

In Section 2, we provide *SHL's formulation, which is mainly motivated by an attempt to minimize the within-group Hamming distances in the code space between a group's codeword and the hash codes of data. With regards to the hash functions, *SHL adopts a kernel-based approach. The aforementioned formulation eventually leads to a minimization problem over the codewords as well as over the Reproducing Kernel Hilbert Space (RKHS) vectors defining the hash functions. A quite noteworthy aspect of the resulting problem is that the minimization over the latter parameters leads to a set of Support Vector Machine (SVM) problems, according to which each SVM generates a single bit of a sample's hash code. In lieu of choosing a fixed, arbitrary kernel function, we use a simple Multiple Kernel Learning (MKL) approach (*e.g.* see [8]) to infer a good kernel from the data. We need to note here that Self-Taught Hashing (STH) [28] also employs SVMs to generate hash codes. However, STH differs significantly from *SHL; its unsupervised and supervised learning stages are completely decoupled, while *SHL uses a single cost function that simultaneously accommodates both of these learning paradigms. Unlike STH, SVMs arise naturally from the problem formulation in *SHL.

Next, in Section 3, an efficient Majorization-Minimization (MM) algorithm is showcased that can be used to optimize *SHL's framework via a Block Coordinate Descent (BCD) approach. The first block optimization amounts to training a set of SVMs, which can be efficiently accomplished by using, for example, LIBSVM [2]. The second block optimization step addresses the MKL parameters, while the third one adjusts the codewords. Both of these steps are computationally fast due to the existence of closed-form solutions.

Finally, in Section 5 we demonstrate the capabilities of *SHL on a series of comparative experiments. The section emphasizes on supervised hash learning problems in the context of CBIR, since the majority of hash learning approaches address this paradigm. We also included some preliminary transductive hash learning results for *SHL as a proof of concept. Remarkably, when compared to other hashing methods on

supervised learning hash tasks, *SHL exhibits the best retrieval accuracy for all the datasets we considered. Some clues to *SHL’s superior performance are provided in Section 4.

2 Formulation

In what follows, $[\cdot]$ denotes the Iverson bracket, *i.e.*, $[\text{predicate}] = 1$, if the predicate is true, and $[\text{predicate}] = 0$, if otherwise. Additionally, vectors and matrices are denoted in boldface. All vectors are considered column vectors and \cdot^T denotes transposition. Also, for any positive integer K , we define $\mathbb{N}_K \triangleq \{1, \dots, K\}$.

Central to hash function learning is the design of functions transforming data to compact binary codes in a Hamming space to fulfill a given machine learning task. Consider the Hamming space $\mathbb{H}^B \triangleq \{-1, 1\}^B$, which implies B -bit hash codes. *SHL addresses multi-class classification tasks with an arbitrary set \mathcal{X} as sample space. It does so by learning a hash function $\mathbf{h} : \mathcal{X} \rightarrow \mathbb{H}^B$ and a set of G labeled codewords $\boldsymbol{\mu}_g$, $g \in \mathbb{N}_G$ (each codeword representing a class), so that the hash code of a labeled sample is mapped close to the codeword corresponding to the sample’s class label; proximity is measured via the Hamming distance. Unlabeled samples are also able to contribute to learning both the hash function and the codewords as it will be demonstrated in the sequel. Finally, a test sample is classified according to the label of the codeword closest to the sample’s hash code.

In *SHL, the hash code for a sample $x \in \mathcal{X}$ is eventually computed as $\mathbf{h}(x) \triangleq \text{sgn } \mathbf{f}(x) \in \mathbb{H}^B$, where the signum function is applied component-wise. Furthermore, $\mathbf{f}(x) \triangleq [f_1(x) \dots f_B(x)]^T$, where $f_b(x) \triangleq \langle w_b, \phi(x) \rangle_{\mathcal{H}_b} + \beta_b$ with $w_b \in \Omega_{w_b} \triangleq \{w_b \in \mathcal{H}_b : \|w_b\|_{\mathcal{H}_b} \leq R_b, R_b > 0\}$ and $\beta_b \in \mathbb{R}$ for all $b \in \mathbb{N}_B$. In the previous definition, \mathcal{H}_b is a RKHS with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_b}$, induced norm $\|w_b\|_{\mathcal{H}_b} \triangleq \sqrt{\langle w_b, w_b \rangle_{\mathcal{H}_b}}$ for all $w_b \in \mathcal{H}_b$, associated feature mapping $\phi_b : \mathcal{X} \rightarrow \mathcal{H}_b$ and reproducing kernel $k_b : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, such that $k_b(x, x') = \langle \phi_b(x), \phi_b(x') \rangle_{\mathcal{H}_b}$ for all $x, x' \in \mathcal{X}$. Instead of a priori selecting the kernel functions k_b , MKL [8] is employed to infer the feature mapping for each bit from the available data. In specific, it is assumed that each RKHS \mathcal{H}_b is formed as the direct sum of M common, pre-specified RKHSs \mathcal{H}_m , *i.e.*, $\mathcal{H}_b = \bigoplus_m \sqrt{\theta_{b,m}} \mathcal{H}_m$, where $\boldsymbol{\theta}_b \triangleq [\theta_{b,1} \dots \theta_{b,M}]^T \in \Omega_{\boldsymbol{\theta}} \triangleq \{\boldsymbol{\theta} \in \mathbb{R}^M : \boldsymbol{\theta} \succeq \mathbf{0}, \|\boldsymbol{\theta}\|_p \leq 1, p \geq 1\}$, \succeq denotes the component-wise \geq relation, $\|\cdot\|_p$ is the usual l_p norm in \mathbb{R}^M and m ranges over \mathbb{N}_M . Note that, if each preselected RKHS \mathcal{H}_m has associated kernel function k_m , then it holds that $k_b(x, x') = \sum_m \theta_{b,m} k_m(x, x')$ for all $x, x' \in \mathcal{X}$.

Now, assume a training set of size N consisting of labeled and unlabeled samples and let \mathcal{N}_L and \mathcal{N}_U be the index sets for these two subsets respectively. Let also l_n for $n \in \mathcal{N}_L$ be the class label of the n^{th} labeled sample. By adjusting its parameters, which are collectively denoted as $\boldsymbol{\omega}$, *SHL attempts to reduce the distortion measure

$$E(\boldsymbol{\omega}) \triangleq \sum_{n \in \mathcal{N}_L} d(\mathbf{h}(x_n), \boldsymbol{\mu}_{l_n}) + \sum_{n \in \mathcal{N}_U} \min_g d(\mathbf{h}(x_n), \boldsymbol{\mu}_g) \quad (1)$$

where d is the Hamming distance defined as $d(\mathbf{h}, \mathbf{h}') \triangleq \sum_b [h_b \neq h'_b]$. However, the distortion E is difficult to directly minimize. As it will be illustrated further below, an upper bound \bar{E} of E will be optimized instead.

In particular, for a hash code produced by *SHL, it holds that $d(\mathbf{h}(x), \boldsymbol{\mu}) = \sum_b [\mu_b f_b(x) < 0]$. If one defines $\bar{d}(\mathbf{f}, \boldsymbol{\mu}) \triangleq \sum_b [1 - \mu_b f_b]_+$, where $[u]_+ \triangleq \max\{0, u\}$ is the hinge function, then $d(\text{sgn } \mathbf{f}, \boldsymbol{\mu}) \leq \bar{d}(\mathbf{f}, \boldsymbol{\mu})$ holds for every $\mathbf{f} \in \mathbb{H}^B$ and any $\boldsymbol{\mu} \in \mathbb{H}^B$. Based on this latter fact, it holds that

$$E(\boldsymbol{\omega}) \leq \bar{E}(\boldsymbol{\omega}) \triangleq \sum_g \sum_n \gamma_{g,n} \bar{d}(\mathbf{f}(x_n), \boldsymbol{\mu}_g) \quad (2)$$

where

$$\gamma_{g,n} \triangleq \begin{cases} [g = l_n] & n \in \mathcal{N}_L \\ [g = \arg \min_{g'} \bar{d}(\mathbf{f}(x_n), \boldsymbol{\mu}_{g'})] & n \in \mathcal{N}_U \end{cases} \quad (3)$$

It turns out that \bar{E} , which constitutes the model’s loss function, can be efficiently minimized by a three-step algorithm, which delineated in the next section.

3 Learning Algorithm

The next proposition allows us to minimize \bar{E} as defined in Equation (2) via a MM approach [7], [6].

Proposition 1. *For any *SHL parameter values ω and ω' , it holds that*

$$\bar{E}(\omega) \leq \bar{E}(\omega|\omega') \triangleq \sum_g \sum_n \gamma'_{g,n} \bar{d}(\mathbf{f}(x_n), \boldsymbol{\mu}_g) \quad (4)$$

where the primed quantities are evaluated on ω' and

$$\gamma'_{g,n} \triangleq \begin{cases} [g = l_n] & n \in \mathcal{N}_L \\ [g = \arg \min_{g'} \bar{d}(\mathbf{f}'(x_n), \boldsymbol{\mu}'_{g'})] & n \in \mathcal{N}_U \end{cases} \quad (5)$$

Additionally, it holds that $\bar{E}(\omega|\omega) = \bar{E}(\omega)$ for any ω . In summa, $\bar{E}(\cdot|\cdot)$ majorizes $\bar{E}(\cdot)$.

Its proof is relative straightforward and is based on the fact that for any value of $\gamma'_{g,n} \in \{0, 1\}$ other than $\gamma_{g,n}$ as defined in Equation (3), the value of $\bar{E}(\omega|\omega')$ can never be less than $\bar{E}(\omega|\omega) = \bar{E}(\omega)$.

The last proposition gives rise to a MM approach, where ω' are the current estimates of the model's parameter values and $\bar{E}(\omega|\omega')$ is minimized with respect to ω to yield improved estimates ω^* , such that $\bar{E}(\omega^*) \leq \bar{E}(\omega')$. This minimization can be achieved via a BCD.

Proposition 2. *Minimizing $\bar{E}(\cdot|\omega')$ with respect to the Hilbert space vectors, the offsets β_p and the MKL weights θ_b , while regarding the codeword parameters as constant, one obtains the following B independent, equivalent problems:*

$$\begin{aligned} \inf_{\substack{w_{b,m} \in \mathcal{H}_m, m \in \mathbb{N}_M \\ \beta_b \in \mathbb{R}, \theta_b \in \Omega_\theta, \mu_{g,b} \in \mathbb{H}}} & C \sum_g \sum_n \gamma'_{g,n} [1 - \mu_{g,b} f_b(x_n)]_+ \\ & + \frac{1}{2} \sum_m \frac{\|w_{b,m}\|_{\mathcal{H}_m}^2}{\theta_{b,m}} \quad b \in \mathbb{N}_B \end{aligned} \quad (6)$$

where $f_b(x) = \sum_m \langle w_{b,m}, \phi_m(x) \rangle_{\mathcal{H}_m} + \beta_b$ and $C > 0$ is a regularization constant.

The proof of this proposition hinges on replacing the (independent) constraints of the Hilbert space vectors with equivalent regularization terms and, finally, performing the substitution $w_{b,m} \leftarrow \sqrt{\theta_{b,m}} w_{b,m}$ as typically done in such MKL formulations (e.g. see [8]). Note that Problem (6) is jointly convex with respect to all variables under consideration and, under closer scrutiny, one may recognize it as a binary MKL SVM training problem, which will become more apparent shortly.

First block minimization: By considering $w_{b,m}$ and β_b for each b as a single block, instead of directly minimizing Problem (6), one can instead maximize the following problem:

Proposition 3. *The dual form of Problem (6) takes the form of*

$$\sup_{\alpha_b \in \Omega_{\alpha_b}} \alpha_b^T \mathbf{1}_{NG} - \frac{1}{2} \alpha_b^T \mathbf{D}_b [(\mathbf{1}_G \mathbf{1}_G^T) \otimes \mathbf{K}_b] \mathbf{D}_b \alpha_b \quad b \in \mathbb{N}_B \quad (7)$$

where $\mathbf{1}_K$ stands for the all ones vector of K elements ($K \in \mathbb{N}$), $\boldsymbol{\mu}_b \triangleq [\mu_{1,b} \dots \mu_{G,b}]^T$, $\mathbf{D}_b \triangleq \text{diag}(\boldsymbol{\mu}_b \otimes \mathbf{1}_N)$, $\mathbf{K}_b \triangleq \sum_m \theta_{b,m} \mathbf{K}_m$, where \mathbf{K}_m is the data's m^{th} kernel matrix, $\Omega_{\alpha_b} \triangleq \{\alpha \in \mathbb{R}^{NG} : \alpha_b^T (\boldsymbol{\mu}_b \otimes \mathbf{1}_N) = 0, \mathbf{0} \preceq \alpha_b \preceq C\boldsymbol{\gamma}'\}$ and $\boldsymbol{\gamma}' \triangleq [\gamma'_{1,1}, \dots, \gamma'_{1,N}, \gamma'_{2,1}, \dots, \gamma'_{G,N}]^T$.

Proof. After eliminating the hinge function in Problem (6) with the help of slack variables $\xi_{g,n}^b$, we obtain the following problem for the first block minimization:

$$\begin{aligned}
& \min_{\substack{w_{b,m}, \beta_b \\ \xi_{g,n}^b}} C \sum_g \sum_n \gamma'_{g,n} \xi_{g,n}^b + \frac{1}{2} \sum_m \frac{\|w_{b,m}\|_{\mathcal{H}_m}^2}{\theta_{b,m}} \\
& \text{s.t. } \xi_{g,n}^b \geq 0 \\
& \quad \xi_{g,n}^b \geq 1 - (\sum_m \langle w_{b,m}, \phi_m(x) \rangle_{\mathcal{H}_m} + \beta_b) \mu_{g,b}
\end{aligned} \tag{8}$$

Due to the Representer Theorem (*e.g.*, see [19]), we have that

$$w_{b,m} = \theta_{b,m} \sum_n \eta_{b,n} \phi_m(x_n) \tag{9}$$

where n is the training sample index. By defining $\boldsymbol{\xi}_b \in \mathbb{R}^{RG}$ to be the vector containing all $\xi_{g,n}^b$'s, $\boldsymbol{\eta}_b \triangleq [\eta_{b,1}, \eta_{b,2}, \dots, \eta_{b,N}]^T \in \mathbb{R}^N$ and $\boldsymbol{\mu}_b \triangleq [\mu_{1,b}, \mu_{2,b}, \dots, \mu_{G,b}]^T \in \mathbb{R}^G$, the vectorized version of Problem (8) in light of Equation (9) becomes

$$\begin{aligned}
& \min_{\boldsymbol{\eta}_b, \boldsymbol{\xi}_b, \beta_b} C \boldsymbol{\gamma}' \boldsymbol{\xi}_b + \frac{1}{2} \boldsymbol{\eta}_b^T \mathbf{K}_b \boldsymbol{\eta}_b \\
& \text{s.t. } \boldsymbol{\xi}_b \succeq \mathbf{0} \\
& \quad \boldsymbol{\xi}_b \succeq \mathbf{1}_{NG} - (\boldsymbol{\mu}_b \otimes \mathbf{K}_b) \boldsymbol{\eta}_b - (\boldsymbol{\mu}_b \otimes \mathbf{1}_N) \beta_b
\end{aligned} \tag{10}$$

where $\boldsymbol{\gamma}'$ and \mathbf{K}_b are defined in Proposition 3. From the previous problem's Lagrangian \mathcal{L} , one obtains

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\xi}_b} = \mathbf{0} \Rightarrow \begin{cases} \boldsymbol{\lambda}_b = C \boldsymbol{\gamma}' - \boldsymbol{\alpha}_b \\ \mathbf{0} \preceq \boldsymbol{\alpha}_b \preceq C \boldsymbol{\gamma}' \end{cases} \tag{11}$$

$$\frac{\partial \mathcal{L}}{\partial \beta_b} = 0 \Rightarrow \boldsymbol{\alpha}_b^T (\boldsymbol{\mu}_b \otimes \mathbf{1}_N) = 0 \tag{12}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\eta}_b} = \mathbf{0} \xrightarrow{\exists \mathbf{K}_b^{-1}} \boldsymbol{\eta}_b = \mathbf{K}_b^{-1} (\boldsymbol{\mu}_b \otimes \mathbf{K}_b)^T \boldsymbol{\alpha}_b \tag{13}$$

where $\boldsymbol{\alpha}_b$ and $\boldsymbol{\lambda}_b$ are the dual variables for the two constraints in Problem (10). Utilizing Equation (11), Equation (12) and Equation (13), the quadratic term of the dual problem becomes

$$\begin{aligned}
& (\boldsymbol{\mu}_b \otimes \mathbf{K}_b) \mathbf{K}_b^{-1} (\boldsymbol{\mu}_b^T \otimes \mathbf{K}_b) = \\
& = (\boldsymbol{\mu}_b \otimes \mathbf{K}_b) (1 \otimes \mathbf{K}_b^{-1}) (\boldsymbol{\mu}_b^T \otimes \mathbf{K}_b) \\
& = (\boldsymbol{\mu}_b \otimes \mathbf{I}_{N \times N}) (\boldsymbol{\mu}_b^T \otimes \mathbf{K}_b) \\
& = (\boldsymbol{\mu}_b \boldsymbol{\mu}_b^T) \otimes \mathbf{K}_b
\end{aligned} \tag{14}$$

Equation (14) can be further manipulated as

$$\begin{aligned}
& (\boldsymbol{\mu}_b \boldsymbol{\mu}_b^T) \otimes \mathbf{K}_b = \\
& = [(\text{diag}(\boldsymbol{\mu}_b) \mathbf{1}_G) (\text{diag}(\boldsymbol{\mu}_b) \mathbf{1}_G)^T] \otimes \mathbf{K}_b \\
& = [\text{diag}(\boldsymbol{\mu}_b) (\mathbf{1}_G \mathbf{1}_G^T) \text{diag}(\boldsymbol{\mu}_b)] \otimes [\mathbf{I}_N \mathbf{K}_b \mathbf{I}_N] \\
& = [\text{diag}(\boldsymbol{\mu}_b) \otimes \mathbf{I}_N] [(\mathbf{1}_G \mathbf{1}_G^T) \otimes \mathbf{K}_b] [\text{diag}(\boldsymbol{\mu}_b) \otimes \mathbf{I}_N] \\
& = [\text{diag}(\boldsymbol{\mu}_b \otimes \mathbf{1}_N)] [(\mathbf{1}_G \mathbf{1}_G^T) \otimes \mathbf{K}_b] [\text{diag}(\boldsymbol{\mu}_b \otimes \mathbf{1}_N)] \\
& = \mathbf{D}_b [(\mathbf{1}_G \mathbf{1}_G^T) \otimes \mathbf{K}_b] \mathbf{D}_b
\end{aligned} \tag{15}$$

Algorithm 1 Optimization of Problem (6)

Input: Bit Length B , Training Samples X containing labeled or unlabeled data.
Output: ω .

1. Initialize ω .
2. **While Not Converged**
3. **For each bit**
4. $\gamma'_{g,n} \leftarrow \text{Equation (5)}$.
5. Step 1: $w_{b,m} \leftarrow \text{Equation (7)}$.
6. $\beta_b \leftarrow \text{Equation (7)}$.
7. Step 2: Compute $\|w_{b,m}\|_{\mathcal{H}_m}^2$.
8. $\theta_{b,m} \leftarrow \text{Equation (16)}$.
9. Step 3: $\mu_{g,b} \leftarrow \text{Equation (17)}$.
10. **End For**
11. **End While**
12. Output ω .

The first equality stems from the identity $\text{diag}(\mathbf{v})\mathbf{1} = \mathbf{v}$ for any vector \mathbf{v} , while the third one stems from the mixed-product property of the Kronecker product. Also, the identity $\text{diag}(\mathbf{v} \otimes \mathbf{1}) = \text{diag}(\mathbf{v}) \otimes \mathbf{I}$ yields the fourth equality. Note that \mathbf{D}_b is defined as in Proposition 3. Taking into account Equation (14) and Equation (15), we reach the dual form stated in Proposition 3. \square

Given that $\gamma'_{g,n} \in \{0, 1\}$, one can easily now recognize that Problem (7) is an SVM training problem, which can be conveniently solved using software packages such as LIBSVM. After solving it, obviously one can compute the quantities $\langle w_{b,m}, \phi_m(x) \rangle_{\mathcal{H}_m}$, β_b and $\|w_{b,m}\|_{\mathcal{H}_m}^2$, which are required in the next step.

Second block minimization: Having optimized over the SVM parameters, one can now optimize the cost function of Problem (6) with respect to the MKL parameters θ_b as a single block using the closed-form solution mentioned in Prop. 2 of [8] for $p > 1$ and which is given next.

$$\theta_{b,m} = \frac{\|w_{b,m}\|_{\mathcal{H}_m}^{\frac{2}{p+1}}}{\left(\sum_{m'} \|w_{b,m'}\|_{\mathcal{H}_{m'}}^{\frac{2p}{p+1}} \right)^{\frac{1}{p}}}, \quad m \in \mathbb{N}_M, b \in \mathbb{N}_B. \quad (16)$$

Third block minimization: Finally, one can now optimize the cost function of Problem (6) with respect to the codewords by mere substitution as shown below.

$$\inf_{\mu_{g,b} \in \mathbb{H}} \sum_n \gamma_{g,n} [1 - \mu_{g,b} f_b(x_n)]_+ \quad g \in \mathbb{N}_G, b \in \mathbb{N}_B \quad (17)$$

On balance, as summarized in Algorithm 1, for each bit, the combined MM/BCD algorithm consists of one SVM optimization step, and two fast steps to optimize the MKL coefficients and codewords respectively. Once all model parameters ω have been computed in this fashion, their values become the current estimate (*i.e.*, $\omega' \leftarrow \omega$), the $\gamma_{g,n}$'s are accordingly updated and the algorithm continues to iterate until convergence is established². Based on LIBSVM, which provides $\mathcal{O}(N^3)$ complexity [12], our algorithm offers the complexity $\mathcal{O}(BN^3)$ per iteration , where B is the code length and N is the number of instances.

4 Insights to Generalization Performance

The superior performance of *SHL over other state-of-the-art hash function learning approaches featured in the next section can be explained to some extend by noticing that *SHL training attempts to minimize

²A MATLAB® implementation of our framework is available at
<https://github.com/yinjiehuang/StarSHL>

the normalized (by B) expected Hamming distance of a labeled sample to the correct codeword, which is demonstrated next. We constrain ourselves to the case, where the training set consists only of labeled samples (*i.e.*, $N = \mathcal{N}_L$, $\mathcal{N}_U = 0$) and, for reasons of convenience, to a single-kernel learning scenario, where each code bit is associated to its own feature space \mathcal{H}_b with corresponding kernel function k_b . Also, due to space limitations, we provide the next result without proof.

Lemma 1. *Let \mathcal{X} be an arbitrary set, $\mathcal{F} \triangleq \{\mathbf{f} : x \mapsto \mathbf{f}(x) \in \mathbb{R}^B, x \in \mathcal{X}\}$, $\Psi : \mathbb{R}^B \rightarrow \mathbb{R}$ be L -Lipschitz continuous w.r.t $\|\cdot\|_1$, then*

$$\hat{\mathfrak{R}}_N(\Psi \circ \mathcal{F}) \leq L \hat{\mathfrak{R}}_N(\|\mathcal{F}\|_1) \quad (18)$$

where \circ stands for function composition, $\hat{\mathfrak{R}}_N(\mathcal{G}) \triangleq \frac{1}{N} \mathbb{E}_{\sigma} \left\{ \sup_{g \in \mathcal{G}} \sum_n \sigma_n g(x_n, l_n) \right\}$ is the empirical Rademacher complexity of a set \mathcal{G} of functions, $\{x_n, l_n\}$ are i.i.d. samples and σ_n are i.i.d random variables taking values with $\Pr\{\sigma_n = \pm 1\} = \frac{1}{2}$.

To show the main theoretical result of our paper with the help of the previous lemma, we will consider the sets of functions

$$\bar{\mathcal{F}} \triangleq \{\mathbf{f} : x \mapsto [f_1(x), \dots, f_B(x)]^T, f_b \in \mathcal{F}_b, b \in \mathbb{N}_B\} \quad (19)$$

$$\begin{aligned} \mathcal{F}_b &\triangleq \{f_b : x \mapsto \langle w_b, \phi_b(x) \rangle_{\mathcal{H}_b} + \beta_b, \beta_b \in \mathbb{R} \text{ s.t. } |\beta_b| \leq M_b, \\ &w_b \in \mathcal{H}_b \text{ s.t. } \|w_b\|_{\mathcal{H}_b} \leq R_b, b \in \mathbb{N}_B\} \end{aligned} \quad (20)$$

Theorem 1. *Assume reproducing kernels of $\{\mathcal{H}_b\}_{b=1}^B$ s.t. $k_b(x, x') \leq r^2$, $\forall x, x' \in \mathcal{X}$. Then for a fixed value of $\rho > 0$, for any $\mathbf{f} \in \bar{\mathcal{F}}$, any $\{\mu_l\}_{l=1}^G$, $\mu_l \in \mathbb{H}^B$ and any $\delta > 0$, with probability $1 - \delta$, it holds that:*

$$er(\mathbf{f}, \mu_l) \leq \hat{er}(\mathbf{f}, \mu_l) + \frac{2r}{\rho B \sqrt{N}} \sum_b R_b + \sqrt{\frac{\log(\frac{1}{\delta})}{2N}} \quad (21)$$

where $er(\mathbf{f}, \mu_l) \triangleq \frac{1}{B} \mathbb{E}\{d(\operatorname{sgn}(\mathbf{f}(x), \mu_l))\}$, $l \in \mathbb{N}_G$ is the true label of $x \in \mathcal{X}$, $\hat{er}(\mathbf{f}, \mu_l) \triangleq \frac{1}{NB} \sum_{n,b} Q_\rho(f_b(x_n) \mu_{l,n,b})$, where $Q_\rho(u) \triangleq \min\left\{1, \max\left\{0, 1 - \frac{u}{\rho}\right\}\right\}$.

Proof. Notice that

$$\begin{aligned} \frac{1}{B} d(\operatorname{sgn}(\mathbf{f}(x), \mu_l)) &= \frac{1}{B} \sum_b [f_b(x) \mu_{l,b} < 0] \leq \frac{1}{B} \sum_b Q_\rho(f_b(x) \mu_{l,b}) \\ &\Rightarrow \mathbb{E} \left\{ \frac{1}{B} d(\operatorname{sgn}(\mathbf{f}(x), \mu_l)) \right\} \leq \mathbb{E} \left\{ \frac{1}{B} \sum_b Q_\rho(f_b(x) \mu_{l,b}) \right\} \end{aligned} \quad (22)$$

Consider the set of functions

$$\Psi \triangleq \{\psi : (x, l) \mapsto \frac{1}{B} \sum_b Q_\rho(f_b(x) \mu_{l,b}), \mathbf{f} \in \bar{\mathcal{F}}, \mu_{l,b} \in \{\pm 1\}, l \in \mathbb{N}_G, b \in \mathbb{N}_B\}$$

Then from Theorem 3.1 of [15] and Equation (22), $\forall \psi \in \Psi$, $\exists \delta > 0$, with probability at least $1 - \delta$, we have:

$$er(\mathbf{f}, \mu_l) \leq \hat{er}(\mathbf{f}, \mu_l) + 2\hat{\mathfrak{R}}_N(\Psi) + \sqrt{\frac{\log(\frac{1}{\delta})}{2N}} \quad (23)$$

where $\hat{\mathfrak{R}}_N(\Psi)$ is the Rademacher complexity of Ψ . From Lemma 1, the following inequality between empirical Rademacher complexities is obtained

$$\hat{\mathfrak{R}}_N(\Psi) \leq \frac{1}{B\rho} \hat{\mathfrak{R}}_N(\|\bar{\mathcal{F}}_\mu\|_1) \quad (24)$$

where $\bar{\mathcal{F}}_\mu \triangleq \{(x, l) \mapsto [f_1(x)\mu_{l,1}, \dots, f_B(x)\mu_{l,B}]^T, f \in \bar{\mathcal{F}} \text{ and } \mu_{l,b} \in \{\pm 1\}\}$. The right side of Equation (24) can be upper-bounded as follows

$$\begin{aligned} \hat{\mathfrak{R}}_N(\|\bar{\mathcal{F}}_\mu\|_1) &= \frac{1}{N} \mathbb{E}_\sigma \left\{ \sup_{\mathbf{f} \in \bar{\mathcal{F}}, \{\mu_{l,n}\} \in \mathbb{H}^B} \sum_n \sigma_n \sum_b |\mu_{l,n,b} f_b(x_n)| \right\} \\ &= \frac{1}{N} \mathbb{E}_\sigma \left\{ \sup_{\mathbf{f} \in \bar{\mathcal{F}}} \sum_n \sigma_n \sum_b |f_b(x_n)| \right\} \\ &= \frac{1}{N} \mathbb{E}_\sigma \left\{ \sup_{\omega_b \in \mathcal{H}_b, \|\omega_b\|_{\mathcal{H}_b} \leq R_b, |\beta_b| \leq M_b} \sum_n \sigma_n \sum_b |\langle w_b, \phi_b(x) \rangle_{\mathcal{H}_b} + \beta_b| \right\} \\ &= \frac{1}{N} \mathbb{E}_\sigma \left\{ \sup_{\omega_b \in \mathcal{H}_b, \|\omega_b\|_{\mathcal{H}_b} \leq R_b, |\beta_b| \leq M_b} \sum_n \sigma_n \sum_b |\langle w_b, \text{sgn}(\beta_b) \phi_b(x) \rangle_{\mathcal{H}_b} + |\beta_b|| \right\} \\ &= \frac{1}{N} \mathbb{E}_\sigma \left\{ \sup_{|\beta_b| \leq M_b} \sum_b [R_b \sqrt{\sigma^T K_b \sigma} + |\beta_b| \sum_n \sigma_n] \right\} \\ &= \frac{1}{N} \mathbb{E}_\sigma \left\{ \sum_b R_b \sqrt{\sigma^T K_b \sigma} \right\} \stackrel{\text{Jensen's Ineq.}}{\leq} \frac{1}{N} \sum_b R_b \sqrt{\mathbb{E}_\sigma \{\sigma^T K_b \sigma\}} \\ &= \frac{1}{N} \sum_b R_b \sqrt{\text{trace}\{K_b\}} \leq \frac{r}{\sqrt{N}} \sum_b R_b \end{aligned} \quad (25)$$

From Equation (24) and Equation (25) we obtain $\hat{\mathfrak{R}}_N(\Psi) \leq \frac{r}{\rho B \sqrt{N}} \sum_b R_b$. Since $\mathfrak{R}_N(\Psi) \triangleq \mathbb{E}_s \{\hat{\mathfrak{R}}_N(\Psi)\}$, where \mathbb{E}_s is the expectation over the samples, we have

$$\mathfrak{R}_N(\Psi) \leq \frac{r}{\rho B \sqrt{N}} \sum_b R_b \quad (26)$$

The final result is obtained by combining Equation (23) and Equation (26). \square

It can be observed that, minimizing the loss function of Problem (6), in essence, also reduces the bound of Equation (21). This tends to cluster same-class hash codes around the correct codeword. Since samples are classified according to the label of the codeword that is closest to the sample's hash code, this process may lead to good recognition rates, especially when the number of samples N is high, in which case the bound becomes tighter.

5 Experiments

5.1 Supervised Hash Learning Results

In this section, we compare *SHL to other state-of-the-art hashing algorithms: Kernel Supervised Learning (KSH) [13], Binary Reconstructive Embedding (BRE) [10], single-layer Anchor Graph Hashing (1-AGH) and its two-layer version (2-AGH) [14], Spectral Hashing (SPH) [26] and Locality-Sensitive Hashing (LSH) [4].

Five datasets were considered: *Pendigits* and *USPS* from the *UCI Repository*, as well as *Mnist*, *PASCAL07* and *CIFAR-10*. For *Pendigits* (10,992 samples, 256 features, 10 classes), we randomly chose 3,000

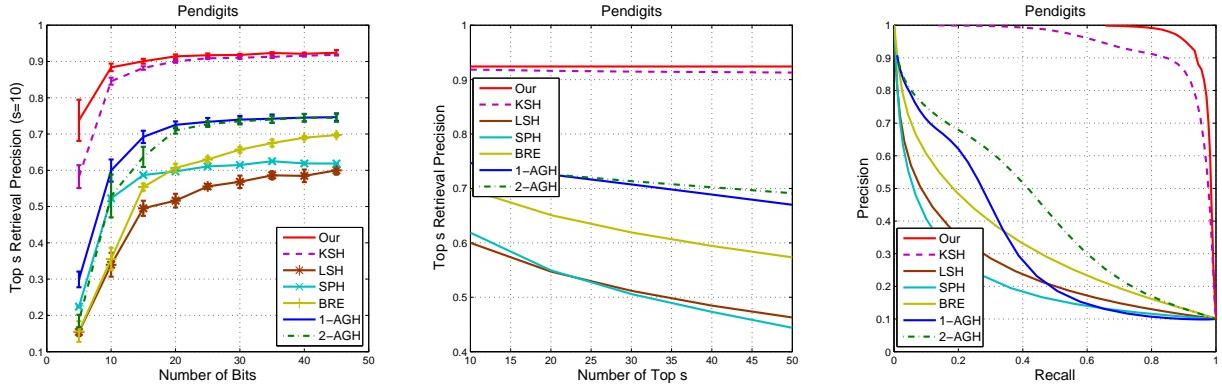


Figure 1: The top s retrieval results and Precision-Recall curve on *Pendigits* dataset over *SHL and 6 other hashing algorithms. (view in color)

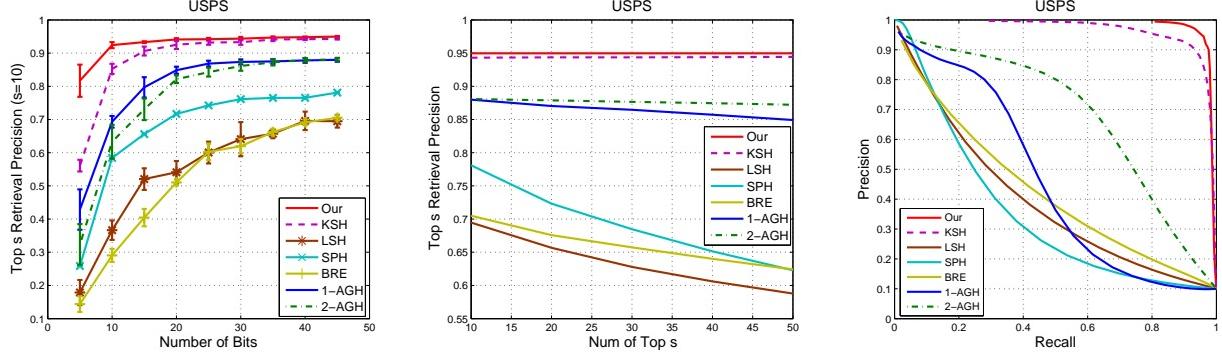


Figure 2: The top s retrieval results and Precision-Recall curve on *USPS* dataset over *SHL and 6 other hashing algorithms. (view in color)

samples for training and the rest for testing; for *USPS* (9,298 samples, 256 features, 10 classes), 3000 were used for training and the remaining for testing; for *Mnist* (70,000 samples, 784 features, 10 classes), 10,000 for training and 60,000 for testing; for *CIFAR-10* (60,000 samples, 1,024 features, 10 classes), 10,000 for training and the rest for testing; finally, for *PASCAL07* (6878 samples, 1,024 features after down-sampling the images, 10 classes), 3,000 for training and the rest for testing.

For all the algorithms used, average performances over 5 runs are reported in terms of the following two criteria: (i) retrieval precision of s -closest hash codes of training samples; we used $s = \{10, 15, \dots, 50\}$. (ii) Precision-Recall (PR) curve, where retrieval precision and recall are computed for hash codes within a Hamming radius of $r \in \mathbb{N}_B$.

The following *SHL settings were used: SVM's parameter C was set to 1000; for MKL, 11 kernels were considered: 1 normalized linear kernel, 1 normalized polynomial kernel and 9 Gaussian kernels. For the polynomial kernel, the bias was set to 1.0 and its degree was chosen as 2. For the bandwidth σ of the Gaussian kernels the following values were used: $[2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}, 1, 2^1, 2^3, 2^5, 2^7]$. Regarding the MKL constraint set, a value of $p = 2$ was chosen. For the remaining approaches, namely KSH, SPH, AGH, BRE, parameter values were used according to recommendations found in their respective references. All obtained results are reported in Figure 1 through Figure 5.

We clearly observe that *SHL performs best among all the algorithms considered. For all the datasets, *SHL achieves the highest top-10 retrieval precision. Especially for the non-digit datasets (*CIFAR-10*,

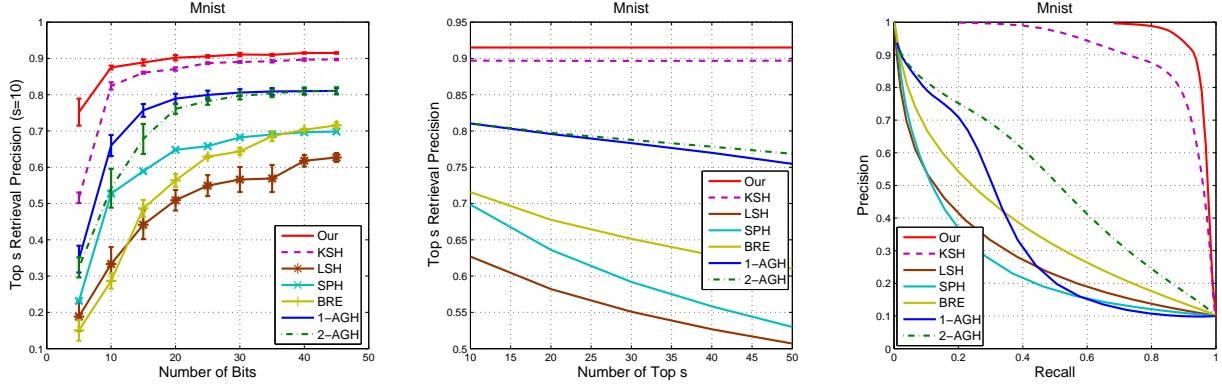


Figure 3: The top s retrieval results and Precision-Recall curve on *Mnist* dataset over *SHL and 6 other hashing algorithms. (view in color)

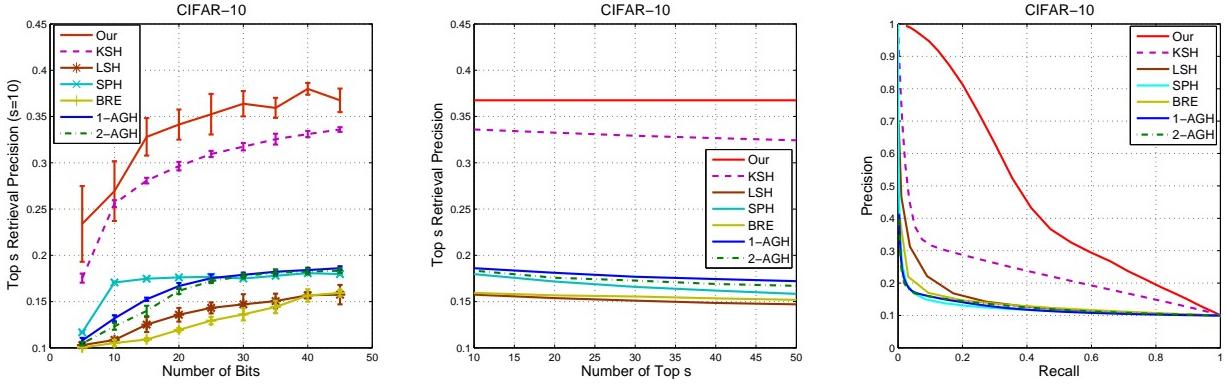


Figure 4: The top s retrieval results and Precision-Recall curve on *CIFAR-10* dataset over *SHL and 6 other hashing algorithms. (view in color)

PASCAL07), *SHL achieves significantly better results. As for the PR-curve, *SHL also yields the largest areas under the curve. Although noteworthy results were reported in [13] for KSH, in our experiments *SHL outperformed it across all datasets. Moreover, we observe that supervised hash learning algorithms, except BRE, perform better than unsupervised variants. BRE may need a longer bit length to achieve better performance as implied by Figure 1 and Figure 3. Additionally, it is worth pointing out that *SHL performed remarkably well for short big lengths across all datasets.

It must be noted that AGH also yielded good results, compared with other unsupervised hashing algorithms, perhaps due to the anchor points it utilizes as side information to generate hash codes. With the exception of *SHL and KSH, the remaining approaches exhibit poor performance for the non-digit datasets we considered.

When varying the top- s number between 10 and 50, once again with the exception of *SHL and KSH, the performance of the remaining approaches deteriorated in terms of top- s retrieval precision. KSH performs slightly worse, when s increases, while *SHL's performance remains robust for *CIFAR-10* and *PASCAL07*. It is worth mentioning that the two-layer AGH exhibits better robustness than its single-layer version for datasets involving images of digits. Finally, Figure 6 shows some qualitative results for the *CIFAR-10* dataset. In conclusion, in our experimentation, *SHL exhibited superior performance for every code length we considered.

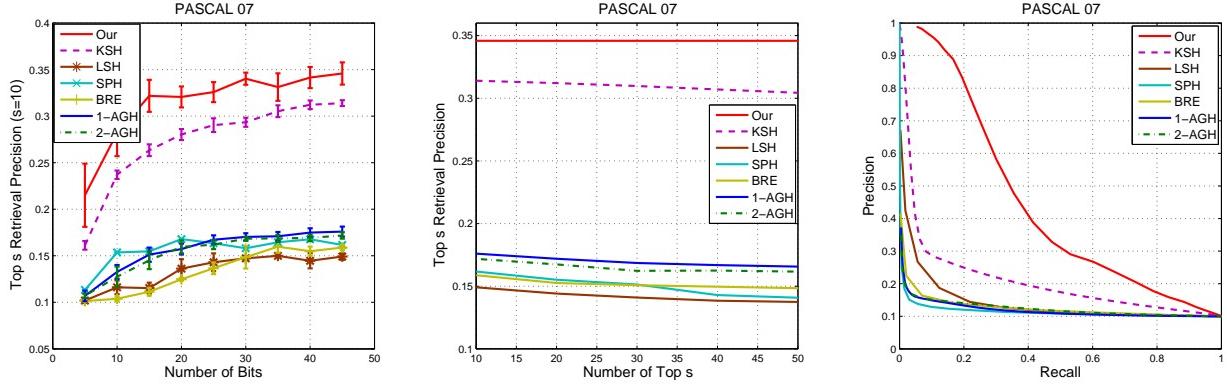


Figure 5: The top s retrieval results and Precision-Recall curve on *PASCAL07* dataset over *SHL and 6 other hashing algorithms. (view in color)

5.2 Transductive Hash Learning Results

As a proof of concept, in this section, we report a performance comparison of our framework, when used in an inductive versus a transductive [23] mode. Note that, to the best of our knowledge, no other hash learning approaches to date accommodate transductive hash learning in a natural manner like *SHL. For illustration purposes, we used the *Vowel* and *Letter* datasets. We randomly chose 330 training and 220 test samples for the *Vowel* and 300 training and 200 test samples for the *Letter*. Each scenario was run 20 times and the code length (B) varied from 4 to 15 bits. The results are shown in Figure 7 and reveal the potential merits of the transductive *SHL learning mode across a range of code lengths.

6 Conclusions

In this paper we considered a novel hash learning framework with two main advantages. First, its Majorization-Minimization (MM)/Block Coordinate Descent (BCD) training algorithm is efficient and simple to implement. Secondly, this framework is able to address supervised, unsupervised and, even, semi-supervised learning tasks in a unified fashion. In order to show the merits of the method, we performed a series of experiments involving 5 benchmark datasets. In these experiments, a comparison between *Supervised Hash Learning (*SHL) to 6 other state-of-the-art hashing methods shows *SHL to be highly competitive.

6.0.1 Acknowledgments

Y. Huang was supported by a Trustee Fellowship provided by the Graduate College of the University of Central Florida. Additionally, M. Georgopoulos acknowledges partial support from NSF grants No. 0806931, No. 0963146, No. 1200566, No. 1161228, and No. 1356233. Finally, G. C. Anagnostopoulos acknowledges partial support from NSF grant No. 1263011. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

- [1] Shumeet Baluja and Michele Covell. Learning to hash: Forgiving hash functions and applications. *Data Mining and Knowledge Discovery*, 17(3):402–430, 2008.
- [2] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

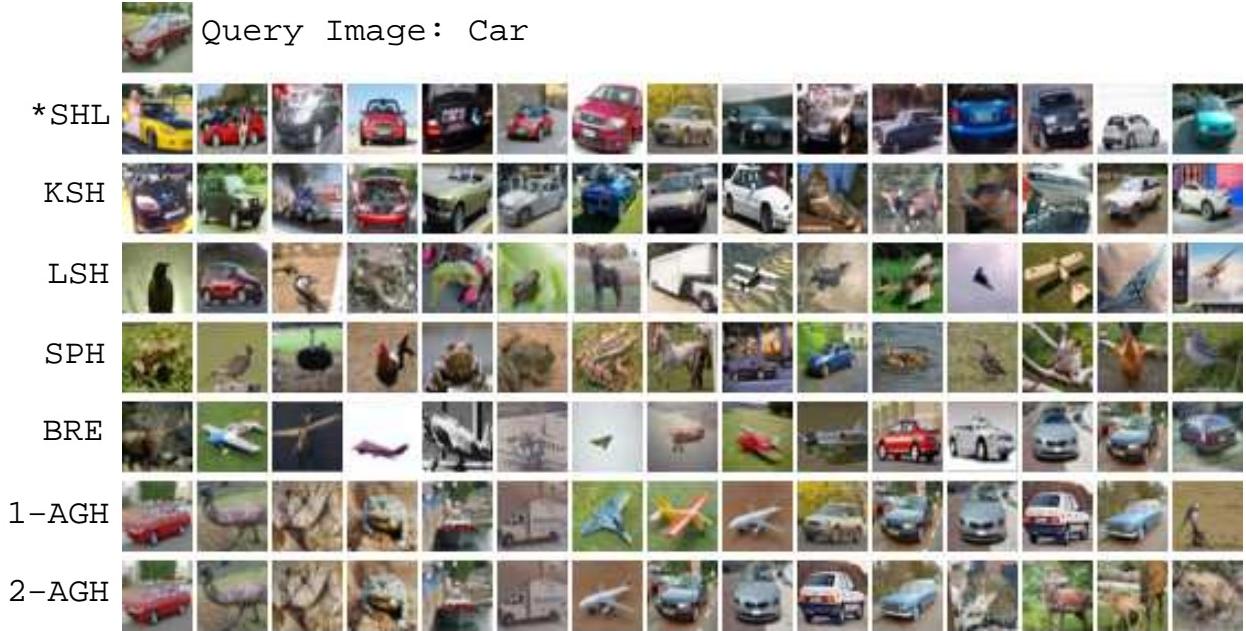


Figure 6: Qualitative results on CIFAR-10. Query image is "Car". The remaining 15 images for each row were retrieved using 45-bit binary codes generated by different hashing algorithms .

- [3] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Ze Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):5:1–5:60, May 2008.
- [4] Aristides Gionis, Piotr Indyk, and Rajeev Motwani. Similarity search in high dimensions via hashing. In *Proceedings of the International Conference on Very Large Data Bases*, pages 518–529, 1999.
- [5] Yunchao Gong and Svetlana Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *Proceedings of Computer Vision and Pattern Recognition*, pages 817–824, 2011.
- [6] David R. Hunter and Kenneth Lange. Quantile regression via an mm algorithm. *Journal of Computational and Graphical Statistics*, 9(1):60–77, Mar 2000.
- [7] David R Hunter and Kenneth Lange. A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37, 2004. [arXiv:<http://dx.doi.org/10.1198/0003130042836>](http://dx.doi.org/10.1198/0003130042836)
- [8] Marius Kloft, Ulf Brefeld, Sören Sonnenburg, and Alexander Zien. l_p -norm multiple kernel learning. *Journal of Machine Learning Research*, 12:953–997, July 2011.
- [9] B. Kulis, P. Jain, and K. Grauman. Fast similarity search for learned metrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2143–2157, 2009.
- [10] Brian Kulis and Trevor Darrell. Learning to hash with binary reconstructive embeddings. In *Proceedings of Advanced Neural Information Processing Systems*, pages 1042–1050, 2009.
- [11] Xi Li, Guosheng Lin, Chunhua Shen, Anton van den Hengel, and Anthony R. Dick. Learning hash functions using column generation. In *Proceedings of the International Conference on Machine Learning*, pages 142–150, 2013.
- [12] Nikolas List and Hans Ulrich Simon. Svm-optimization and steepest-descent line search. In *Proceedings of the Conference on Computational Learning Theory*, 2009.

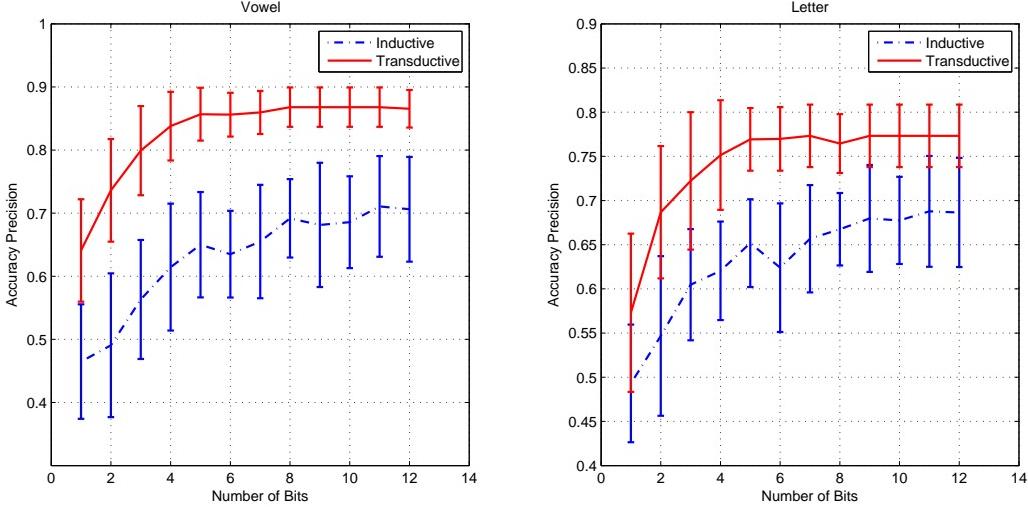


Figure 7: Accuracy results between Inductive and Transductive Learning.

- [13] Wei Liu, Jun Wang, Rongrong Ji, Yu-Gang Jiang, and Shih-Fu Chang. Supervised hashing with kernels. In *Proceedings of Computer Vision and Pattern Recognition*, pages 2074–2081, 2012.
- [14] Wei Liu, Jun Wang, Sanjiv Kumar, and Shih-Fu Chang. Hashing with graphs. In *Proceedings of the International Conference on Machine Learning*, pages 1–8, 2011.
- [15] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012.
- [16] Yadong Mu, Jialie Shen, and Shuicheng Yan. Weakly-supervised hashing in kernel space. In *Proceedings of Computer Vision and Pattern Recognition*, pages 3344–3351, 2010.
- [17] Mohammad Norouzi and David J. Fleet. Minimal loss hashing for compact binary codes. In *Proceedings of the International Conference on Machine Learning*, pages 353–360, 2011.
- [18] Ruslan Salakhutdinov and Geoffrey Hinton. Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978, July 2009.
- [19] Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *Proceedings of the European Conference on Computational Learning Theory*, pages 416–426, 2001.
- [20] Gregory Shakhnarovich, Paul Viola, and Trevor Darrell. Fast pose estimation with parameter-sensitive hashing. In *Proceedings of the International Conference on Computer Vision*, pages 750–, 2003.
- [21] Christoph Strecha, Alex Bronstein, Michael Bronstein, and Pascal Fua. Ldahash: Improved matching with smaller descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):66–78, January 2012.
- [22] A. Torralba, R. Fergus, and Y. Weiss. Small codes and large image databases for recognition. In *Proceedings of Computer Vision and Pattern Recognition*, pages 1–8, 2008.
- [23] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.
- [24] Jun Wang, S. Kumar, and Shih-Fu Chang. Semi-supervised hashing for large-scale search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(12):2393–2406, 2012.

- [25] Jun Wang, Sanjiv Kumar, and Shih-Fu Chang. Sequential projection learning for hashing with compact codes. In *Proceedings of the International Conference on Machine Learning*, pages 1127–1134, 2010.
- [26] Yair Weiss, Antonio Torralba, and Robert Fergus. Spectral hashing. In *Proceedings of Advanced Neural Information Processing Systems*, pages 1753–1760, 2008.
- [27] Chun-Nam John Yu and Thorsten Joachims. Learning structural svms with latent variables. In *Proceedings of the International Conference on Machine Learning*, pages 1169–1176, 2009.
- [28] Dell Zhang, Jun Wang, Deng Cai, and Jinsong Lu. Self-taught hashing for fast similarity search. In *Proceedings of the International Conference on Research and Development in Information Retrieval*, pages 18–25, 2010.